Multiple description video coding through adaptive segmentation

Brian A. Heng and Jae S. Lim

Research Laboratory of Electronics
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ABSTRACT

Multiple description video coding is one method that can be used to reduce detrimental effects caused by transmission over lossy packet networks. In a multiple description system, a video sequence is segmented into two or more complimentary streams in such a way that each stream is independently decodable. When combined, the streams provide the highest level of quality, yet if one of the streams is lost or delivered late the video can be played out with only a slight reduction in overall quality. Each approach to multiple description coding consists of a tradeoff between compression efficiency and robustness. How efficiently each method achieves this tradeoff depends on the level of quality and robustness desired and on the characteristics of the video itself. Previous approaches to multiple description coding have made the assumption that a single segmentation method would be used for an entire sequence. Yet, the optimal method of segmentation can vary depending on the goals of the system, it can change over time, and it can vary within a frame. This work introduces a unique approach to multiple description coding through the use of adaptive segmentation. By selecting from a set of segmentation methods, the system adapts to the local characteristics of the video and maximizes tradeoff efficiency. We present an overview of this system and analyze its performance on real video sequences.

Keywords: video coding, multiple description, adaptive segmentation, error-resilient, packet video

1. INTRODUCTION

Streaming video applications have become increasingly popular over the past few years, and by all indications their use will continue to grow in the future. However, the Internet provides only a best-effort service, characterized by variable bandwidths, packet losses and delays. This environment is inhospitable to real-time applications which require a minimum quality of service in order to maintain synchronization between the sender and receiver. Applications must be able to withstand changing conditions on the network or they can suffer severe performance degradations. For many applications, these problems can be solved to some extent with a suitable amount of buffering at the receiver. However, buffering introduces an additional delay into the system which is unacceptable for interactive applications.

There are many examples of interactive video applications including video conferencing and video-on-demand. Both of these require a high degree of communication between opposite ends of the network and stringent demands on end-to-end delay. In the case of video conferencing, there exists a limit on the amount of delay which can exist between two users attempting to maintain a reasonable conversation. Once this limit is exceeded, the two parties can no longer interact without significant effort. Thus, significant buffering is not an option. Even in applications where some amount of buffering is acceptable, the amount of buffering necessary in any situation is unknown ahead of time due to the time-varying properties of the network. Occasionally network links fail altogether, and there may be some extended period of time during which two points in the network cannot talk to one another at all. This type of outage can underflow any reasonably sized buffer. For this reason, current approaches to interactive video streaming often suffer from severe glitches each time the network becomes congested.

E-mail: heng@mit.edu; jslim@mit.edu
This work has been sponsored in part by Hewlett-Packard Research Laboratories, Palo Alto, CA. The views expressed are those of the authors and may not represent the views of the Hewlett-Packard Company.

Multiple description (MD) video coding is one method that can be used to reduce the detrimental effects caused by this type of best-effort network. In a multiple description system, a video sequence is segmented into two or more complementary streams in such a way that each stream is independently decodable. When combined, the streams provide the highest level of quality, but even independently they are able to provide an acceptable level of quality. These streams can then be sent along separate paths through the network to experience more or less independent losses and delays. In the event that a portion of one of the streams is lost or delivered late, the video playback will not suffer a severe glitch or stop completely to allow for rebuffering. On the contrary, the remaining stream(s) will continue to be played out with only a slight reduction in overall quality.

Previous approaches to multiple description coding have made the assumption that one approach must be used for an entire sequence. Yet, the optimal method of segmentation will depend on many factors including the amount of motion in the scene, the amount of spatial detail, desired levels of quality, current network conditions, and so on. Each segmentation method proposed consists of a tradeoff between compression efficiency and robustness. How efficiently each method achieves this tradeoff depends on the quality of video desired, the preferred level of robustness, and the characteristics of the video itself. Some methods work well when looking for slight increases in robustness but are very inefficient when pushed to their limits. Other methods are well suited for adding a significant amount of redundancy but require too much overhead when absolute reliability is not necessary. Some methods work well in stationary regions, others work well in moving regions. Thus, it is not an optimal choice to use any one segmentation method for an entire sequence.

This paper investigates the use of multiple modes of segmentation within a given sequence. In this way the MD encoder adapts to local characteristics of the video as well as to selected rate-distortion goals of the system. This paper introduces the proposed adaptive segmentation system and presents performance results in comparison to the corresponding nonadaptive systems. Section 2 provides a brief introduction to multiple description coding. Sections 3 and 4 present an overview of the adaptive segmentation approach and discuss the details of the system which has been implemented based on this concept. Experimental results are provided in Section 5 which demonstrate the potential of the adaptive segmentation system. A summary of this work is provided in Section 6.

2. MULITPLE DESCRIPTION CODING

This paper focuses on improving streaming video applications through the use of multiple description coding. A multiple description coder segments a stream into two or more separately decodable streams and transmits these independently over the network. The quality of the received video improves with each received description, but the loss of any one of these streams does not cause complete failure. Thus video playback can continue, at a slight reduction in quality, without waiting for rebuffering or retransmission.

As a simple example, consider a MD system in which a sequence is segmented in two by spatially sub-sampling each frame. For instance, the even numbered lines of each frame could be placed in one sequence and the odd numbered lines placed in another. This process generates two new sequences with half the vertical resolution. These two sequences can then be independently coded and transmitted across the network. In the event that both streams are received, the lines from each can be interleaved to reconstruct the full resolution sequence. In the event that one stream is lost, any number of interpolation techniques could be used to estimate the missing lines. In general, the distortion resulting from this estimation will be higher than that obtained when both streams are received, yet playback can continue despite the total loss of one stream.

Of course, this gain in robustness comes with a cost. Spatially sub-sampling each frame lowers the spatial correlation, thus reducing coding efficiency and increasing the number of bits necessary to maintain the same level of quality. The total bit rate necessary for this MD system to achieve a given distortion will in general be higher than the corresponding rate for a single stream encoder to achieve the same distortion. It is a tradeoff between coding efficiency and robustness. However, in the type of application under consideration, it is not so much a question of whether it is useful to give up some amount of efficiency for an increase in reliability as it is a question of finding the most effective way to achieve this tradeoff.

Many approaches have been previously suggested for multiple description coding. Some of the many contributions include, multiple description quantization[1], correlating transforms[2], spatial segmentation[3], transform domain segmentation[4], and temporal segmentation[5]. For an in depth review of MD coding see the overview by Goyal[6].

In the case of two independent streams, the problem can be formally defined as follows, (see Figure 1). A given source is encoded into two separate streams at rate R_1 and R_2. These two streams are independently sent over two channels through the network. The receiver consists of three separate sub-decoders, the central decoder which outputs the sequence when both streams arrive intact, and two side decoders which output the sequence when only one of the streams has arrived. The distortion in the output from the central decoder is defined as D_0, and the two side distortions are defined as D_1 and D_2. The state selector outputs the video stream from the most appropriate decoder.

The goal is to find an optimal quintuple, $\{R_1, R_2, D_0, D_1, D_2\}$ for a given situation. The relative importance of each of these five factors depends on the situation at hand. In some systems, reliability may be most important in which case D_1 and D_2 become more significant and the importance of bit rate perhaps decreases. In some cases it is the opposite, and robustness may be sacrificed for coding efficiency. The system does not need to be balanced either; R_1 could be weighted more heavily than R_2, and so on.

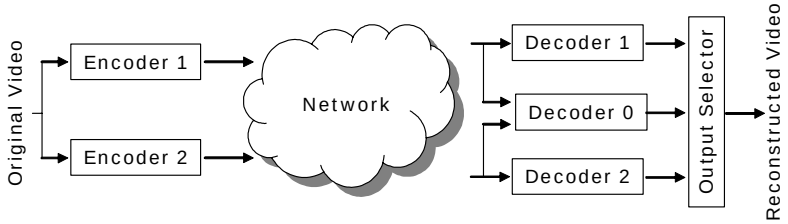


Figure 1. Model of a two stream multiple description encoder-decoder system. The input is compressed using two complementary encoders. Each of these streams is sent independently through the network. Depending on which streams are correctly received, the receiver decodes one of the two independent streams or combines the results of both.

It should be noted here that multiple description coding is not the same as scalable video coding. Similar to MD coding, a scalable coder encodes a sequence into multiple streams called layers, see Figure 2. However, scalable coding makes use of a single independent base layer followed by one or more dependent enhancement layers. This allows some receivers to decode only the base layer to receive basic video, while others can decode the base layer and one or more enhancement layers to achieve improved quality, spatial resolution, and/or frame rate. Unlike MD coding, the loss of the base layer renders the enhancement layer(s) useless. In some sense, scalable coding is a special case approach to multiple description coding where it is assumed that the base layer will be delivered with absolute reliability.

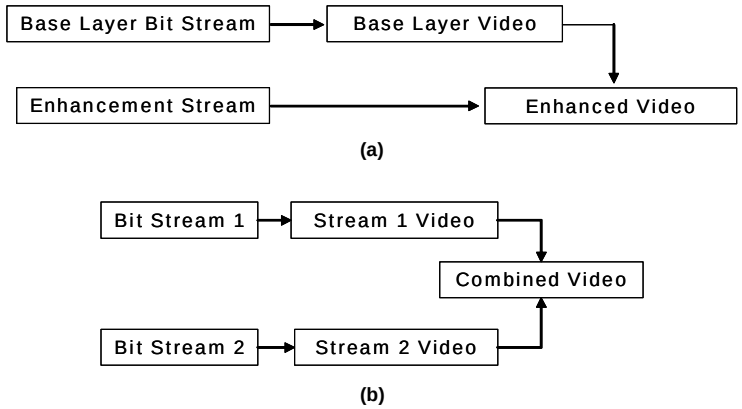


Figure 2. An example of scalable video coding versus multiple description coding. In scalable coding (a) the enhancement layer(s) are dependent on the base layer while in multiple description coding (b), each stream is independently coded.

3. ADAPTIVE SEGMENTATION

In order to improve compression efficiency, video compression systems attempt to remove as much correlation as possible from a video sequence prior to coding. However, each approach to MD coding creates two or more descriptions with some amount of correlation between them. This correlation helps to improve the performance of the system in the event one of the descriptions must be estimated from the remaining descriptions, yet it reduces the compression efficiency of the video coder. How efficiently each method achieves this tradeoff depends on the quality of video desired, the preferred level of robustness, and the characteristics of the video itself. Most approaches to MD coding are done in a nonadaptive fashion where one method is used for the entire sequence. However, since the encoder in this system has access to the original sequence, it could easily measure the resulting rate-distortion statistics and could adaptively select a method for a given region in an intelligent manner

The encoder in the adaptive segmentation system presented in this paper analyzes local regions of the input sequence, and selects a segmentation mode for that particular region. The encoder has full knowledge of the reconstruction methods used by the decoder in any of the possible loss-scenarios, and thus, it can fully calculate the rate-distortion set $\{R_1, R_2, D_0, D_1, D_2\}$ during encoding and choose the best method for the current block. It can choose from among the given set the single method which meets bit rate, distortion, and robustness goals most efficiently. The end result is a system that adapts to both the characteristics of the video as well as the performance goals of the system.

The focus of this paper is on a balanced two channel system, where both channels are considered of equal importance. This is a special case of the more general MD problem, but it is common in practice since the time varying properties of the network make it difficult to characterize or prioritize any path. In this case the problem formulation simplifies to optimizing a three parameter set, the total bit rate $R_{Total} = R_1 + R_2$, the central distortion $D_{Central} = D_0$, and the average side channel distortion $D_{Side} = \text{average}(D_1, D_2)$. The system could be generalized to handle more than two channels or a situation in which one channel takes preference over the other in a fairly straightforward manner.

The goal of this system is to minimize some function of the central and side distortions subject to a bit rate constraint.

$$\text{minimize } f\left(D_{Central}, D_{Side}\right) \qquad \text{given the constraint } \mathrm{R}_{\mathrm{Total}} \le R \tag{1}$$

Define $R_{i,j}$ as the number of bits necessary to code region i with segmentation mode j, and $D_{Central\,i,j}$ and $D_{Side\,i,j}$ as the resulting distortions. This goal of adaptive segmentation may then be summarized as follows. For a given total rate *R*, choose the set of modes, j, which satisfy $\sum_i R_{i,j} \le R$ and minimize the distortion metric $\sum_i f\left(D_{Central}, D_{Side}\right)$.

The function $f\left(D_{Central}, D_{Side}\right)$ used in the above approach could, in practice, be any reasonable function. However, it should be chosen in such a way that the minimization of this function attempts to maximize the utility of the end user. For instance, one reasonable approach is to attempt to minimize the expected distortion. Under certain assumptions about the loss characteristics of the network, an expression for the expected distortion in the output can be obtained. By using this expression as the distortion metric, $f\left(D_{Central}, D_{Side}\right)$, the above approach will minimize the expected distortion of the system.

One simple function that could be used for this purpose is a weighted average

$$f\left(D_{Central}, D_{Side}\right) = (1-\alpha)\cdot D_{Central} + \alpha \cdot D_{Side} \qquad 0 \le \alpha \le 1 \tag{2}$$

This approach in a sense measures the tradeoff efficiency of each mode, where α defines the relative value of robustness versus coding efficiency. The parameter α can be adjusted based on the performance goals of the system and expected loss characteristics of the network. For example, in a system where coding efficiency is the primary goal, α can be set to 0. This would amount to choosing the set of modes which minimize central distortion. Similarly, if robustness happened to be the most significant factor, α could be set to 1, which would result in choosing those modes which minimize the side distortion. The parameter α in effect defines an acceptable tradeoff ratio between the two distortions.

It should be noted here that it is in general not possible to simultaneously minimize both the central and side distortions. For a given mode, fixing any one parameter in the set $\{R_{Total}, D_{Central}, D_{Side}\}$ completely specifies the remaining two parameters. For example, consider two extreme modes, a multiple description repetition coder repeating all data in both

streams versus a standard single stream encoder. For a given rate, the single stream encoder will be far more efficient in general and would have a much lower distortion than the repetition approach. At the same time, the excessive redundancy of the repetition coder would result in a significantly lower side distortion. Minimizing the central distortion would require the encoder to choose only the single stream mode, while minimizing the side distortion would result in choosing only the repetition mode. For this reason, it becomes necessary to minimize a function of the central and side distortions as in the weighted approach suggested above.

4. SYSTEM IMPLEMENTATION

The system presented below has been used to examine two test sequences (Foreman and News). Both of these test sequences consisted of 50 frames of progressive scan video at CIF resolution (288 rows x 352 columns) in 4:2:0 YUV format. For the purposes of this paper, the distortion caused by losses during data compression as well as losses during network transmission have been quantitatively measured using PSNR (the peak-signal-to-noise ratio of the luminance component between the original and reconstructed video) and bit rates have been expressed in bits-per-pixel (BPP). It should be noted that PSNR and perceived quality are not always directly correlated. Higher PSNR does not necessarily indicate higher quality video, but the use of PSNR is a common practice and has been found to be a useful estimate of video quality. All three components of the original 4:2:0 YUV sequences have been coded to calculate bit rates despite the fact that only the luminance component is used in PSNR analysis.

The adaptive segmentation MD coder described in the following sections has been implemented based on the main profile of the H.264/MPEG4 Part 10 advanced video coding standard[7]. The current implementation uses a single I-frame (intra-coded frame) followed by P-frames (inter-predicted frames) and refreshes the prediction loop by periodically updating lines of macroblocks using intra-frame coding. Each line of macroblocks is placed into a slice and packetized using RTP (real-time transport protocol) by the H.264 encoder. The following sections describe the functionality of the adaptive system in further detail.

4.1. Frame Partitioning

Adaptive segmentation is performed by partitioning each frame into blocks of 16x16 pixels and selecting one segmentation method for each of these blocks. The most appropriate block size used for segmentation remains an area of interest since smaller block sizes and/or adaptive block sizing could improve the performance of the system but has not been investigated in this paper. As the blocks are made smaller, the system can adapt more efficiently to the local regions of the image. However, using smaller blocks increases the number of blocks per frame, which will lead to more overhead when the mode for each block is encoded. At some point the benefits of finer adaptation may be overshadowed by the increase in bit rate.

4.2. Segmentation Mode Set

There are a number of methods previously developed to solve the MD problem, each with their own strengths and weaknesses. For adaptive segmentation to be effective, it is useful to find a set which complements each other well; one method strong where another is weak and vice versa. Some well-rounded methods that provide reasonable performance under any situation may not turn out to be as useful in this system as methods which provide excellent results in their own specialized cases. For this reason, particular attention should be given to methods which provide exceptional performance in certain situations (moving regions, stationary regions, etc…), but may have been previously overlooked due to having poor performance when applied to an entire video sequence. It is also necessary to choose an appropriate number of modes. Additional modes require additional overhead since the decoder will need to know which mode was selected. Assuming the number of available modes is small, the overhead required to code each mode is relatively minor, perhaps only 1-2 bits per block. However, as the number of modes increases this overhead could become significant.

Selecting a set of useful methods is certainly an interesting problem, yet the optimization of this set and of the methods themselves is not the main focus of this work. Note that it is always possible to add improved methods to the system in

the future with expectations of similar gains in performance, as long as both the encoder and decoder are updated accordingly.

The system described in this paper uses three possible segmentation modes; temporal segmentation, spatial segmentation, or no segmentation at all. The no-segmentation method essentially performs standard single description encoding; all data for the current block is placed in one description or the other. This method is the most efficient method in terms of reducing bit rates, however if a block is lost there is no information available in the opposite stream to help in reconstruction. In this case the system simply repeats the data from the previously decoded frame. In many regions, (e.g. stationary areas) this type of reconstruction is good enough, so there is no need to sacrifice coding efficiency in an attempt to improve error resilience.

The temporal segmentation method used in the adaptive system separates the original sequence along the temporal direction. Even frames are predicted only from even frames and odd frames are predicted only from odd frames. This is essentially the same as the video redundancy coding (VRC) method which is a portion of the H.263 standard for providing error resilience[8]. This approach creates two separate prediction loops, where a loss of data in an even frame does not propagate to the odd frames, and vice versa. The fact that the neighboring frames remain unaffected by a loss allows for more sophisticated reconstruction methods as has been suggested by Apostolopoulos[5]. The current system makes use of a motion compensated interpolation scheme similar to that presented by Wong and Au[9]. Temporal segmentation is less efficient than single stream encoding since temporal prediction efficiency in general decreases as the temporal distance between frames increases.

The spatial segmentation method used is essentially the spatial equivalent of the temporal method presented above. Here the encoder splits the data into even and odd lines rather than frames. Even lines are predicted only from previous even lines and odd lines are predicted only from previous odd lines. The even data is placed into one description and the odd data is placed into the other. In the event of a loss, half of the lines may be correctly reconstructed using the lines contained in the opposite description, and the missing lines may be reconstructed using standard interpolation techniques. This method tends to complement the temporal method well. It tends to have higher performance in the regions where translational block motion estimation fails to accurately model the scene.

4.3. Optimal Mode Selection

Given the problem formulation above, the optimal mode choice can be determined using Lagrangian optimization techniques[10]. As stated previously, the objective of the adaptive segmentation system is to solve the budget constrained allocation problem presented in Section 3.

The goal is to minimize the distortion metric,

$$\sum_i f\left(D_{Central_{i,j}}, D_{Side_{i,j}}\right), \tag{3}$$

while satisfying the bit rate constraint

$$\sum_i R_{i,j} \le R. \tag{4}$$

This problem can be solved using the discrete version of Lagrangian optimization. Rather than attempting to solve the above budget constrained minimization directly, a Lagrangian cost function can be minimized instead.

$$J(\lambda) = \sum_i \left(f\left(D_{Central_{i,j}}, D_{Side_{i,j}}\right) + \lambda R_{i,j} \right) \tag{5}$$

The Lagrange multiplier λ is a non-negative real number which can be varied to achieve various operational points. When this Lagrangian cost function is minimized, if λ is chosen such that the budget constraint is satisfied, the distortion metric will be minimized as well.

Thus, the optimal solution to the given problem can be determined by encoding the current block with each segmentation mode and choosing the mode which minimizes the Lagrangian cost function. The value of λ can be varied until the given budget constraint is satisfied. For instance, setting $\lambda = 0$ is equivalent to minimizing the

distortion metric without a budget constraint, while setting λ arbitrarily high is equivalent to minimizing the bit rate. Once the budget constraint has been met, the resulting mode decisions will achieve the minimum distortion.

The one exception to this approach comes when considering the temporal segmentation method. The only difference between standard single description coding and temporal segmentation is whether the current block is predicted from the previous frame or from two frames ago. This reference picture choice certainly affects the coding efficiency of the current block, yet it has little effect on the reconstruction quality in the event this block is lost. If the current block is lost, all prediction information is lost as well, so it makes little difference if it had been predicted from one or two frames ago. Consider the situation where the previous frame has been lost. If the current block is predicted from two frames ago, it will be unaffected by this loss and could be used effectively for temporal interpolation of the missing frame. However, if the current block is predicted from the missing frame, it is no longer available to assist in reconstruction and some other approach like frame repetition must be used. Thus, the choice of reference frame can potentially have a significant effect on the reconstruction quality of the *previous* frame. This additional distortion in the previous frame due to the reference picture choice can be calculated during encoding and can be compared against the additional bit rate necessary to predict from two frames ago, and thus the reference picture choice can also be optimized in the Lagrangian rate-distortion manner presented above.

4.4. Overview of Macroblock Encoding Process

An overview of proposed encoding process based on these concepts is presented below. Figure 3 presents a flowchart of the macroblock encoding process. The system codes each block using each mode and then reconstructs the blocks under three possible situations; no streams are lost, stream #1 is lost, and stream #2 is lost. The resulting distortions for these three scenarios are calculated and the weighted average of the three is generated as suggested in (2). This weighted average and the corresponding number of bits required for this mode are used to calculate the Lagrangian cost function. Once all modes have finished, the mode which minimizes the cost function is selected and the bits for this block are written into the output stream.

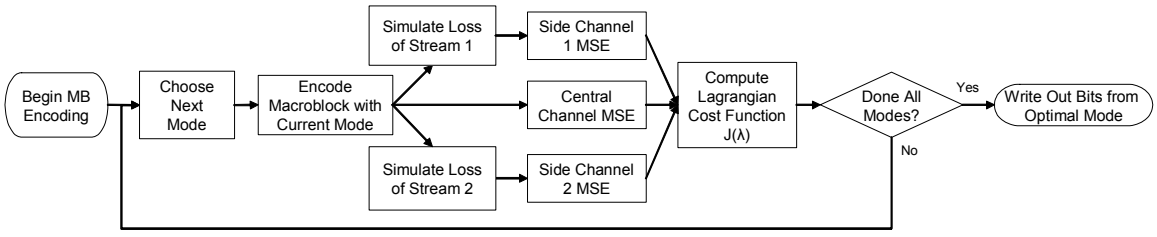


Figure 3. Flowchart demonstrating the proposed macroblock encoding process. Each macroblock (MB) is coded with each mode and the corresponding central and side distortions are calculated. Lagrangian optimization is used to choose which mode to write to the output stream.

5. EXPERIMENTAL RESULTS

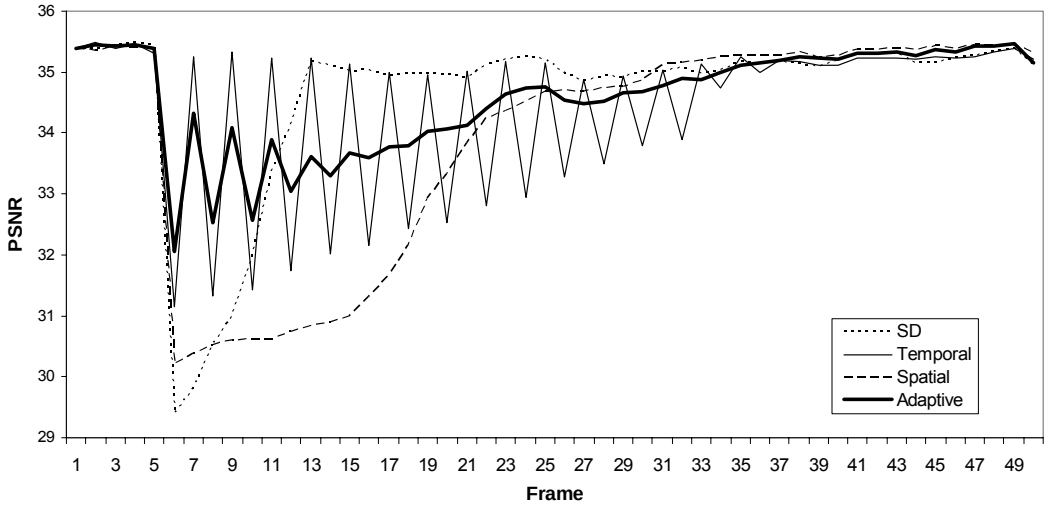
The results of coding the Foreman and News sequences with the adaptive segmentation system are presented in the following sections. In each case the adaptive system is compared against its nonadaptive counterparts. The weighted average distortion from (2) was used with $\alpha = 0.4$. Section 5.1 shows the loss-recovery characteristics of each of the methods under a single loss event. Section 5.2 shows the rate-distortion curves for each of the systems under a simulated binomial packet loss model where each packet has a 5% probability of being lost independent of other packets.

5.1. Single Loss Event

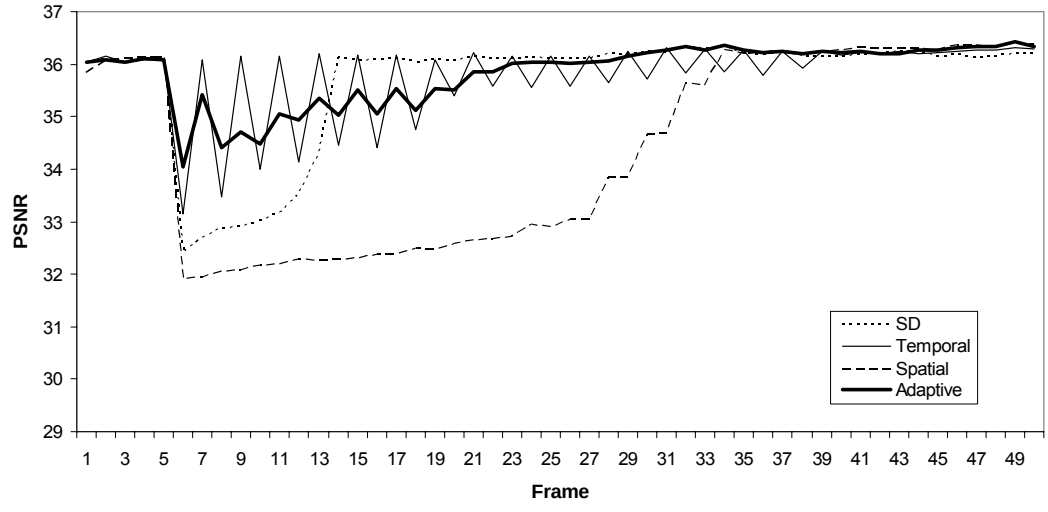
This section illustrates the performance of each of the systems after a single frame has been lost. The decoder simulated a loss of the sixth frame to demonstrate the severity of this loss for each of the systems and to show the recovery from

this error. The results of this experiment are shown in Figure 4. The four systems shown here are single description coding (SD), temporal segmentation, spatial segmentation, and adaptive segmentation.

In order to make a fair comparison, the systems have been adjusted to have the same central distortion. This distortion was held approximately constant across all frames at around 35.5 dB for the Foreman sequence and 36dB for the News sequence. Each of the systems has varying levels of efficiency/redundancy and thus would in general require different bit rates to maintain the same central distortion. To account for this fact while still maintaining a fair comparison, the frequency of intra-coding has been adjusted in each system until the bit rates of all the systems were as close as possible to identical (about .25 BPP for the Foreman sequence and .18 BPP for the News sequence). In general more frequent intra-coding helps to increase the speed at which a system recovers from an error, yet requires a higher bit rate. The excess bit rate available in the more efficient systems is used to increase the rate of recovery from the loss.



(a)



(b)

Figure 4. Loss/Recovery characteristics after single frame loss. (a) Forman sequence. (b) News sequence. The decoder simulated the loss of the sixth frame to analyze the severity of this loss for each of the systems and to demonstrate how each system recovers from this error.

The SD coding method has the highest coding efficiency and thus recovers fastest from the error due to more frequent intra coding. It could be argued that this quick recovery indicates the SD system has the highest performance. However, the initial error due to this loss tends to be quite severe and very objectionable, while the remaining systems are able to mitigate this initial error more effectively and more smoothly recover from the loss. The SD coding method performs slightly better in the News sequence since this sequence contains a large stationary background region which is reconstructed nearly perfectly by the frame repetition approach. However, the regions which contain motion are reconstructed poorly with the SD method. This type of large localized error is something not well captured by PSNR analysis.

The temporal segmentation approach significantly improves the reconstruction of the missing frame compared to the SD method. An improvement of approximately 2dB is seen in frame 6 of the Foreman sequence. This improvement is not as significant in the News sequence, again due to the larger stationary background region. Due to the fact that the even frames and odd frames have separate prediction loops in the temporal approach, the odd frames are unaffected by the loss of the sixth frame. This fact helps to improve the recovery of the missing frame, yet this oscillation between high and low distortion can also create objectionable flicker in the resulting sequence. As was previously mentioned, the motion compensated prediction used in the temporal method will in general be less efficient than the SD method, so the system uses less frequent intra coding than the SD method and recovers from the error at a slower rate.

Similar to the temporal method, the spatial segmentation system attempts to reduce the initial error due to the loss at the cost of having a slower recovery time. Compared to the temporal method, the reconstruction quality of the spatial method tends to be slightly lower and the recovery rate is about the same or maybe slightly faster. In the News sequence, the spatial method is actually the worst performer in the PSNR sense, although one can argue that perceptually, the large localized errors in the SD method are more objectionable than the smaller distributed errors in the spatial approach.

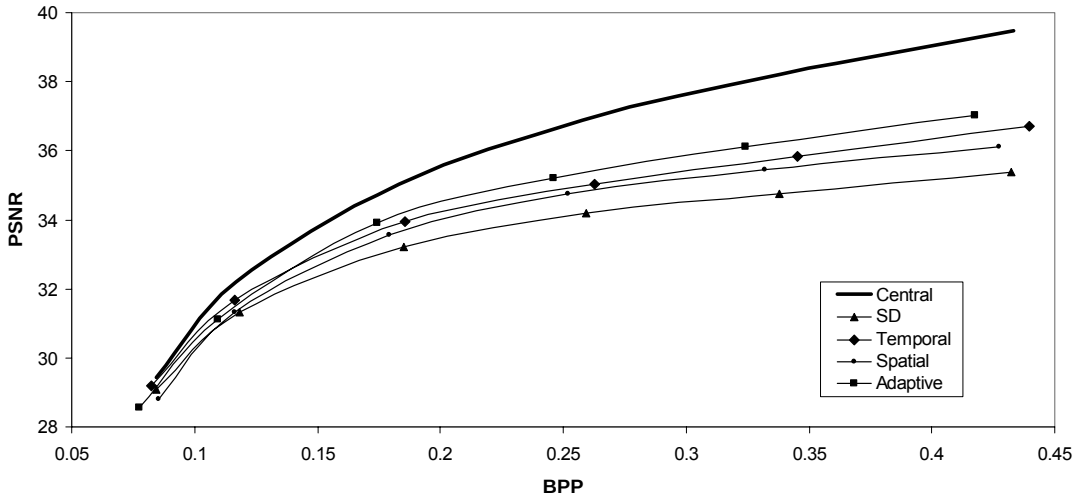
The adaptive system is able to take advantage of the strengths of each of the systems and in both sequences demonstrates the highest performance of the group. It is able to reduce the initial error caused by the loss more effectively than any of the non-adaptive methods and recovers from this error as fast as or faster than either the spatial or temporal methods. The adaptive system is making decisions to give up the coding efficiency of the SD method in those blocks where it determines the gain in robustness from the spatial or temporal methods outweighs the corresponding loss in coding efficiency.

5.2. Rate-Distortion Curves under Binomial Loss Model

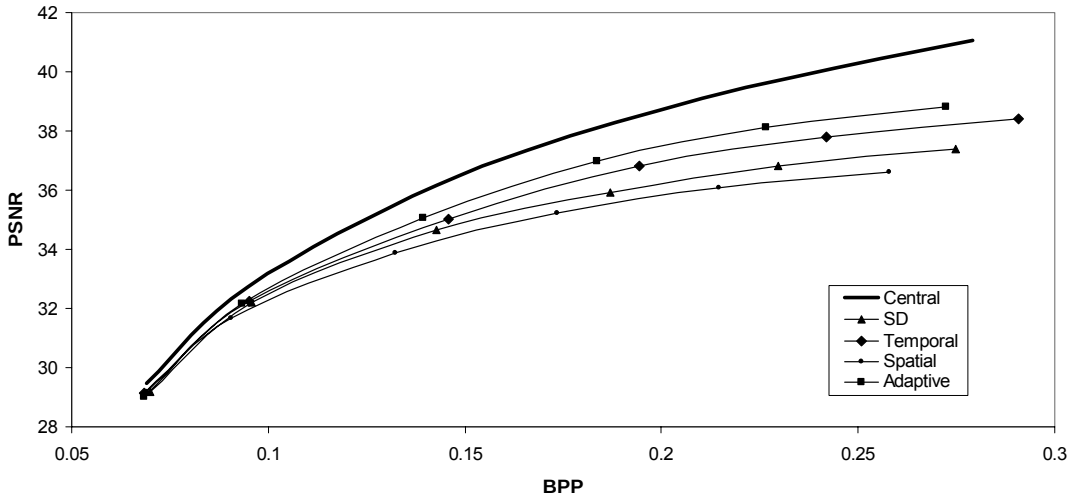
While the analysis above is useful for demonstrating the loss/recovery characteristics of each system, it does not provide a very realistic loss scenario and fails to give a true representation of system performance. A single frame loss does not model actual packet loss on a network and the results can vary significantly depending on which frame was lost. To account for this, a similar experiment was run with two main differences. First, rather than simulating the loss of a whole frame, the following experiment was run by simulating a binomial packet loss model. In this model, each packet has a fixed probability of being lost independent of other packets; specifically the packet loss rate used was 5%. Secondly, the PSNR results have been averaged over all 50 frames and the experiment has been performed at various bit rates to generate rate-distortion (RD) curves for each system. These RD curves can be found in Figure 5.

As was done in section 5.1, the intra-coding frequency of each system has been adjusted such that each system achieved the same distortion under loss-free conditions while simultaneously requiring the same bit rate. Thus all systems have identical central distortion curves. This curve is represented by the bold line in Figure 5. Each of the remaining lines in Figure 5 represents the RD curve for one of the systems after 5% packet loss.

When analyzing this result, two things should be pointed out. First, as was mentioned in section 5.1, the News sequence has a large stationary background region which is reconstructed nearly perfectly by either the SD or temporal methods. The large localized errors which do show up may be very visible perceptually while not having a significant effect on the PSNR. Thus, these methods may have a slight advantage when considering only PSNR. Secondly, as was demonstrated in Figure 4, the temporal method has a significant amount of oscillation between frames with low distortion and frames with high distortion. When averaged across all frames, the perceptual flicker this generates is not taken into account.



(a)



(b)

Figure 5. Rate-distortion curves under 5% binomial packet loss. (a) Forman sequence. (b) News sequence. The bold central distortion line represents the equal performance of all the systems when no data is lost. The remaining lines represent the performance of each of the individual systems after losses have occurred.

In both sequences, the adaptive system is able to increase the quality of the resulting video over the nonadaptive methods at a fixed bit rate or reduce the bit rate at a fixed level of quality. This gain becomes more significant at higher bit rates as the overhead required for adaptive segmentation has less of an effect.

CONCLUSION

This paper has introduced the concept of adaptive segmentation for multiple description coding. The system proposed makes use of multiple modes of segmentation within a given sequence allowing it to adapt to local characteristics of the video as well as to selected rate-distortion goals of the system. One such system has been presented in some detail along with performance results which have shown the potential for this adaptive approach to significantly improve video

quality. The effectiveness of this adaptive scheme will certainly depend on the source material, and the results from the two sequences presented above cannot possibly hope to accurately represent the entire collection of possible video sequences. Even so, the results are quite promising, and it is apparent that the benefits of adaptive segmentation can be significant. Even the combination of a small number of simple segmentation methods can result in a considerable increase in error robustness.

REFERENCES

1. V. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Transactions on Information Theory* **39**, pp. 821-834, May 1993.
2. M. Orchard, Y. Wang, V. Vaishampayan, and A. Reibman, "Redundancy rate-distortion analysis of multiple description coding using pairwise correlating transforms," *Proc. of the IEEE Intl. Conference on Image Processing* **1**, pp. 608-611, October 1997.
3. V. Vaishampayan and J. Domaszewicz, "Design of entropy-constrained multiple-description scalar quantizers," *IEEE Transactions on Information Theory* **40**, pp. 245-250, January1994.
4. I. Bajic and J. Woods, "Domain-based multiple description coding of images and video," *IEEE Transactions on Image Processing* **12**, pp. 1211-1225, October 2003.
5. J. Apostolopoulos, "Error-resilient video compression via multiple state streams," in *Proc. of the IEEE Intl. Conference on Image Processing* **3**, pp. 352-355, September 2000.
6. V. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Processing Magazine* **18**, pp. 74-93, September 2001.
7. ITU-T Rec. H.264/ISO/IEC, "Advanced video coding for generic audiovisual services", 2003.
8. S. Wenger, "Video redundancy coding in H.263+," in *Proc. of the Intl. Workshop on Audio-Visual Services over Packet Networks*, Aberdeen, U.K., September 1997.
9. C. Wong and O. Au, "Fast motion compensated temporal interpolation for video" in *Proc. SPIE: Visual Communication and Image Processing* **2501**, pp. 1108-1118, April 1995.
10. G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine* **15**, pp. 74-90, November 1998.